

Identifying Causal Variants by Fine Mapping Across Multiple Studies

Nathan LaPierre^{1,†,*}, Kodi Taraszka^{1,†}, Helen Huang², Rosemary He³,
Farhad Hormozdiari⁶, and Eleazar Eskin^{1,4,5}

¹ Department of Computer Science, University of California, Los Angeles, CA, 90095, United States

² Department of Ecology and Evolutionary Biology, Los Angeles, CA, 90095, United States

³ Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, 90095, United States

⁴ Department of Human Genetics, University of California, Los Angeles, CA, 90095, United States

⁵ Department of Computational Medicine, University of California, Los Angeles, CA, 90095, United States

⁶ Harvard T.H. Chan School of Public Health, Boston, MA 02115, United States

† These authors contributed equally

* Email corresponding author at: nlapier2@cs.ucla.edu

Abstract

Increasingly large Genome-Wide Association Studies (GWAS) have yielded numerous variants associated with many complex traits, motivating the development of “fine mapping” methods to identify which of the associated variants are causal. Additionally, GWAS of the same trait for different populations are increasingly available, raising the possibility of refining fine mapping results further by leveraging different linkage disequilibrium (LD) structures across studies. Here, we introduce multiple study causal variants identification in associated regions (MsCAVIAR), a method that extends the popular CAVIAR fine mapping framework to a multiple study setting using a random effects model. MsCAVIAR only requires summary statistics and LD as input, accounts for uncertainty in association statistics using a multivariate normal model, allows for multiple causal variants at a locus, and explicitly models the possibility of different SNP effect sizes in different populations. In a trans-ethnic, trans-biobank Type 2 Diabetes analysis, we show that MsCAVIAR returns causal set sizes that are over 20% smaller than those given by current state of the art methods for trans-ethnic fine-mapping.

Introduction

Genome-Wide Association Studies (GWAS) have successfully identified numerous genetic variants associated with a variety of complex traits in humans [1–3]. However, most of these associated variants are not causal, and are simply in Linkage Disequilibrium (LD) with the true causal variants. Identifying these causal variants is a crucial step towards understanding the genetic architecture of complex traits, but testing all associated variants at each locus using functional studies is cost-prohibitive. This problem is addressed by statistical “fine mapping” methods, which attempt to prioritize a small subset of variants for further testing while accounting for LD structure [4].

The classic approach to fine mapping involves simply selecting a given number of SNPs with the strongest association statistics for follow-up, but this performs sub-optimally because it does not account for LD structure [5]. Bayesian methods that did account for LD structure were developed [6, 7], but were based upon the simplifying assumption that each locus only harbors a single causal variant, which is not true in many cases [8]. Additionally, many early methods required individual-level genetic data, whereas many human GWAS often provide only summary statistics due to privacy concerns. CAVIAR [8] introduced a Bayesian approach that relied only on summary statistics and LD, accounted for uncertainty in association statistics using a multivariate normal (MVN) distribution, and allowed for the possibility of multiple causal SNPs at a locus. This approach was widely adopted and later made more efficient by CAVIARBF [9] and FINEMAP [10].

There is growing interest in improving fine-mapping by leveraging information from multiple studies. One of the most important examples of this is trans-ethnic fine mapping, which can significantly improve fine mapping power and resolution by leveraging the distinct LD structures in each population, as seen in methods such as trans-ethnic PAINTOR [11] and MR-MEGA [12]. Intuitively, the set of SNPs that are tightly correlated with the causal SNP(s) will be different in different populations, allowing more SNPs to be filtered out as potential candidates. However, the varying LD patterns also present a unique challenge in the multiple study setting that trans-ethnic fine mapping methods must handle. Additionally, while there is evidence that the same SNPs drive association signals across populations, there is also heterogeneity in their effect sizes, presenting another challenge [13]. Existing methods either assume a single causal SNP at each locus [12, 14] or do not explicitly model heterogeneity [11], limiting their power [15].

In this paper, we present MsCAVIAR, a novel method that addresses these challenges. We retain the Bayesian MVN framework of CAVIAR while introducing a novel approach to explicitly account for the heterogeneity of effect sizes between studies using a Random-Effects (RE) model. Our method requires only summary statistics and LD matrices as input, allows for multiple causal variants at a locus, and models uncertainty in association statistics and between-study heterogeneity. The output is a set of SNPs that, with a user-set confidence threshold (e.g. 95%), contains all causal SNPs at the locus.

We show in simulation studies that MsCAVIAR outperforms existing trans-ethnic fine mapping methods [11] and extensions of methods such as CAVIAR [8] to the multiple study setting. In a trans-ethnic, trans-biobank analysis of Type 2 Diabetes, we demonstrate that MsCAVIAR significantly improves the resolution of fine mapping compared to trans-ethnic PAINTOR or running CAVIAR on either population individually. MsCAVIAR is freely available at <https://github.com/nlapier2/MsCAVIAR>.

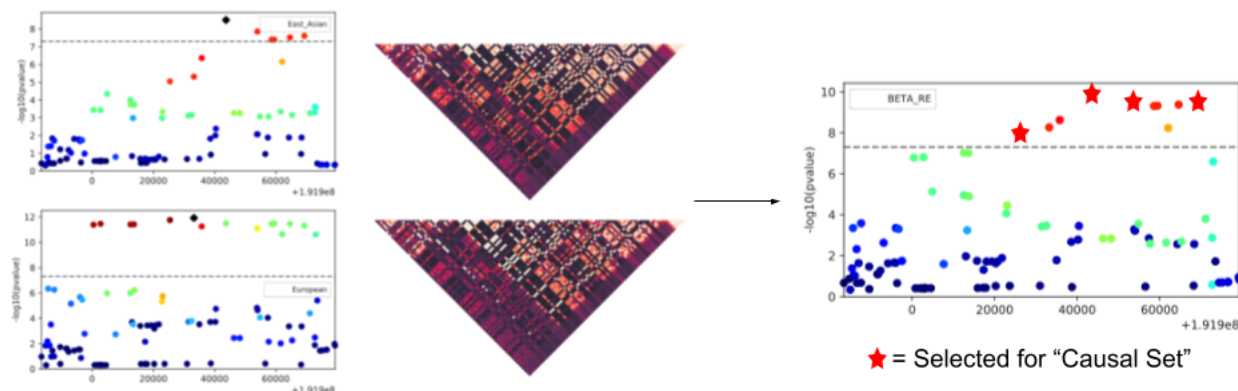


Fig. 1. Overview of MsCAVIAR. MsCAVIAR takes as input the Z-scores and LD matrices for SNPs at a locus in two or more studies (left). Based on this input, MsCAVIAR leverages the different LD structures and models SNP effect heterogeneity to produce a refined “causal set” of SNPs (shown as red stars above) for follow-up functional validation studies. This set is often smaller than the set of SNPs that are significant in meta-analysis results (right).

Results

MsCAVIAR overview

Our method, MsCAVIAR, takes as input the association statistics (e.g. Z-scores) for SNPs at the same locus in multiple studies and the linkage disequilibrium (LD) structure between variants obtained from in-sample genotyped data. MsCAVIAR computes and outputs a minimal-sized “causal set” of SNPs that, with probability at least ρ , contains *all* causal SNPs. This process is visualized in Figure 1.

By our definition of a causal set, every causal SNP must be contained in the set with high probability, but not every SNP in the set need to be causal. Concretely, each SNP can be assigned a binary causal status: 1 for causal or 0 for non-causal. So long as none of the SNPs outside of the causal set are set to 1, the assignments are compatible with our definition of a causal set. We can represent these causal status assignments in a binary vector with one entry for each SNP denoting its causal status; we call such a vector a “configuration” and denote it as C . For each configuration C compatible with the causal set, we compute its (posterior) probability in a Bayesian manner: the probability of a configuration of SNPs being causal given the association statistics can be computed by modeling a prior probability for that configuration and a likelihood function for the association statistics given the assumed causal SNPs given by C (see Methods for details).

The overall likelihood function can be decomposed into a product over the likelihood function for each study, since we assume that the studies are independent. More specifically, we assume that there is a true global effect size for a SNP over all possible populations, around which the effect sizes for that SNP in different studies are independently drawn according to a heterogeneity variance parameter (Methods). This allows MsCAVIAR to model the fact that effect sizes of a SNP across different studies are related, but not equal. Because we expect the summary statistics to be a function of their LD with the causal SNPs, the parameters of the likelihood function for each study are different, assuming the studies have different LD patterns. By computing the product over the likelihood of each study, we are able to account for their different LD patterns in determining the likelihood over all the studies.

The posterior probability for a causal set is then computed by summing the posterior probabilities of all compatible configurations, and then dividing by the sum of the posterior probabilities for all possible configurations. We start by assessing causal sets containing only one SNP, and continue

increasing the size of the causal sets analyzed until one of them exceeds the posterior probability threshold ρ . In practice, ρ is set to a high value such as 95%.

MsCAVIAR improves fine mapping resolution in a simulation study

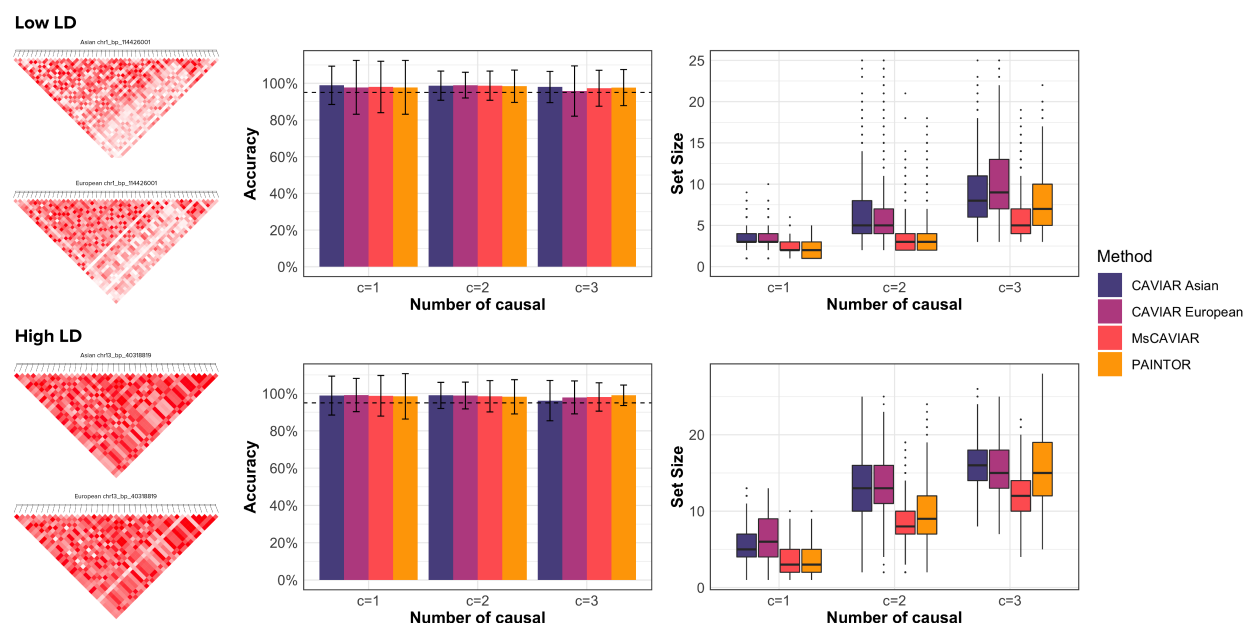


Fig. 2. Comparison of accuracy and set size using simulated data. We simulated a trans-ethnic GWAS by using LD matrices generated from European and East Asian populations in the 1000 Genomes project. One relatively low LD region and one relatively high LD region were chosen (far left). Using these LD matrices, we implanted either 1, 2, or 3 causal SNPs and simulated their effect sizes. For each number of causal SNPs, we performed 1000 simulations (e.g. re-picking the causal SNPs and re-drawing the causal SNP effect sizes). In this figure, we report the average accuracy and standard deviation for each method in the bar graph (middle) and the set size in the box-plot (right). All methods were run with posterior probability threshold $\rho^* = 0.95$, so methods with 95% or higher accuracy were considered “well-calibrated” (dashed line in the bar plots).

In order to evaluate the performance of MsCAVIAR as compared with other methods, we performed a simulation study. In order to select realistic loci for fine-mapping, we identified regions in a trans-ethnic GWAS of rheumatoid arthritis [16] that contained peak SNPs with p-values of less than 0.0001 and contained ten or more SNPs in a 100kbp region centered around that peak. For each such locus, we used the 1000 Genomes project [17] to generate LD matrices for the SNPs at that locus for both European and East Asian populations. Out of these loci, we selected one region with relatively low LD, where 20% of the SNPs have LD equal to or higher than 0.5, and one region with relatively high LD, where 80% of the SNPs have LD equal to or higher than 0.5 (Figure 2, LD matrices). These represent easier and more difficult scenarios, respectively, for fine mapping, since LD makes signals more difficult to distinguish. We pruned groups of SNPs that were in perfect LD in one or more of the populations, leaving one SNP for each. If a group of SNPs were in perfect LD in one population, but not the other, we retained the SNP with the highest Z-score in the other population in order to retain the most signal.

Using these LD matrices, we implanted causal SNPs and simulated their effect sizes. In each simulation, we implanted either 1, 2, or 3 causal SNPs. Each causal SNP’s true non-centrality

parameter λ was drawn according to $\mathcal{N}(5.2, 0.125^2)$. We then drew the non-centrality parameter for each study i according to $\lambda_i \sim \mathcal{N}(\lambda, 0.5)$, and subsequently the summary statistics according to $S_i \sim \mathcal{N}(\lambda_i \Sigma_i, \Sigma_i)$. For each number of causal SNPs, we performed 1000 replicate simulations (e.g. re-drawing the causal SNP effect sizes and re-picking the causal SNPs).

Using this data, we compared MsCAVIAR to the trans-ethnic mode of PAINTOR [11] and to CAVIAR [8] run on East Asians and Europeans, individually (Figure 2). All methods were run with posterior probability threshold $\rho^* = 0.95$, so methods with 95% or higher accuracy were considered “well-calibrated” (dashed line in the bar plots). MsCAVIAR’s heterogeneity parameter was set to $\tau^2 = 0.5$ (Methods); different settings for τ^2 gave similar results. All methods were well-calibrated in both LD settings (Figure 2, bar plots). This is unsurprising since CAVIAR is well-known to be calibrated in the single study setting [8–10], as is PAINTOR in the trans-ethnic setting [11].

However, when considering the subset of simulations in which each method was able to correctly capture all causal variants (i.e. 100% accuracy), we observed that MsCAVIAR consistently returns the smallest average set size (Figure 2, box plots). MsCAVIAR and PAINTOR return smaller set sizes than CAVIAR run on either population across all settings, highlighting the value of using varying LD patterns in different populations to refine fine-mapping results. MsCAVIAR returned smaller set sizes than PAINTOR with multiple causal variants or high LD. This may be due to MsCAVIAR’s explicit modeling of heterogeneity between studies. In both the high LD and multiple causal variants setting, complex and strong correlations between non-causal and causal SNPs are induced, and modeling heterogeneity between studies allows for more effective use of the differing LD structures to disentangle non-causal from causal SNPs.

MsCAVIAR is well-calibrated with different population sizes between studies

It is possible that input studies can have different sample sizes, in which case the effect sizes of their SNPs is expected to be different proportionally to sample size, in addition to heterogeneity. We tested whether MsCAVIAR would still be well-calibrated in this setting, and compared it again with trans-ethnic PAINTOR and with CAVIAR run on the individual populations (Figure 3).

In order to evaluate performance under this scenario in a simulation study, we used the same LD matrices from the previous section, but now varied the population size for one of the studies. We fixed the population size of the Asian study at 10,000 individuals, and varied the European study to have population sizes of 1, 2, 5, or 10 times that of the Asian study. Consequently, the effect sizes of causal SNPs in the European study were larger than those of the corresponding SNPs in the Asian study by a factor of $\sqrt{1}$, $\sqrt{2}$, $\sqrt{5}$, and $\sqrt{10}$ (Methods). For the sake of sufficient statistical power, we ensured that the causal variants in the smaller study were still statistically significant genome-wide. 1000 simulation replicates were run for each LD setting. In each simulation, we implanted three causal SNPs and simulated their effect sizes, with the association statistics of non-causal SNPs being based on their correlation with causal SNPs (Methods). All methods were run with posterior probability threshold $\rho^* = 0.95$, so methods with 95% or higher accuracy were considered “well-calibrated” (dashed line in the bar plots). MsCAVIAR was run with its heterogeneity parameter set at $\tau^2 = 0.5$ (Methods).

Once again, MsCAVIAR was well-calibrated and generally returned the smallest causal set sizes. As the sample size difference grew, the difference between MsCAVIAR, CAVIAR on Europeans, and PAINTOR tended to diminish. This is likely due to the fact that we required SNPs to be genome-wide significant in the smaller study, such that the larger study had very large effect sizes for causal SNPs when there was a significant sample size imbalance, making the fine mapping problem easier. Reinforcing this interpretation is the fact that CAVIAR on Asians had consistently much larger causal set sizes than the other methods when the sample size imbalance was large.

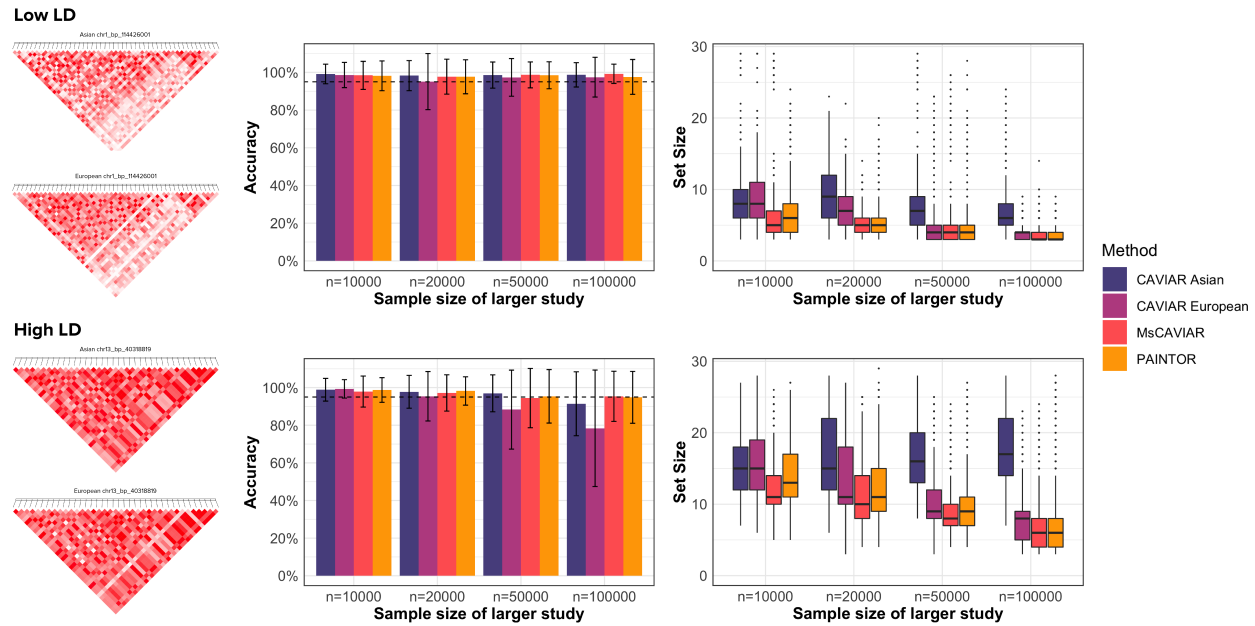


Fig. 3. Comparison of accuracy and set size using simulated studies with unequal sample sizes. We simulated a trans-ethnic GWAS by using LD matrices generated from European and East Asian populations in the 1000 Genomes project. Using these LD matrices, we implanted 3 causal SNPs and simulated their effect sizes. In each set of simulations, we fix the population size of the Asian study at 10,000 individuals, and vary the European study to have population sizes of 1, 2, 5, or 10 times that of the Asian study. In this figure, we report the average accuracy and standard deviation for each method in the bar graph (left) and the set size in the box-plot (right).

All methods were well-calibrated in the low LD setting, but we observed that as the sample size increases with high LD that CAVIAR's calibration on the larger population decreases. This is likely due to the extremity of the situation, with exceptionally large effect sizes in combination with the high LD setting.

MsCAVIAR improves fine mapping resolution in trans-ethnic Type 2 Diabetes analysis

In order to evaluate the performance of MsCAVIAR on real data, we performed a trans-ethnic, trans-biobank fine mapping analysis of Type 2 Diabetes (T2D) using summary statistics from the UK Biobank (UKB) [18] and Biobank Japan (BBJ) [19] projects. These studies involved 361,194 and 191,764 people, respectively. Only White Europeans from the UK Biobank were used. To generate loci for fine mapping, we centered 100kbp windows around genome wide-significant peak SNPs (p -value $\leq 5 \times 10^{-8}$), discarding all SNPs with p -values above 0.0001, as they were highly unlikely to be informative. We applied several other filters to identify loci where trans-ethnic fine mapping was worthwhile. If a SNP was genome wide-significant in one ethnic GWAS but had a p -value above 0.0001 in the other ethnicity, we did not allow this to be a peak SNP, as information from the other ethnicity would be unlikely to help improve resolution in this case. In these instances, fine mapping within one population would be sufficient. We also excluded all loci with fewer than ten SNPs with p -values below 0.0001 in each study, as fine mapping is not as useful when there are few strongly associated SNPs. We excluded loci from chromosome six, where there were numerous statistically significant SNP effect sizes due to the presence of human leukocyte antigen (HLA) regions.

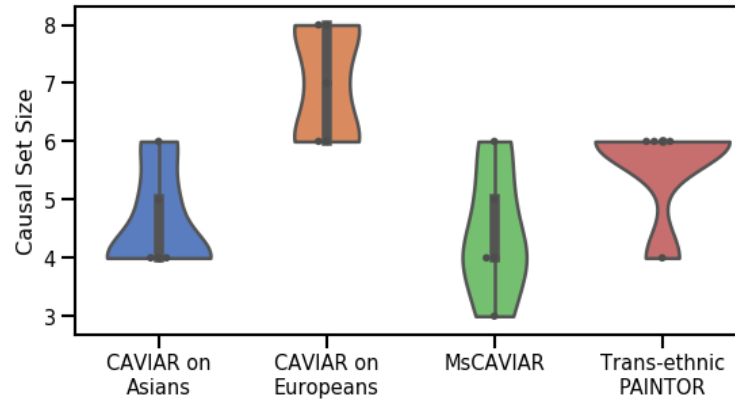


Fig. 4. MsCAVIAR improves fine mapping resolution in trans-ethnic Type 2 Diabetes analysis. We compare the results of MsCAVIAR when applied to two Type 2 Diabetes (T2D) GWAS, White European people from the UK Biobank [18] and Japanese people from the Japan Biobank [19], versus trans-ethnic PAINTOR [11] and applying CAVIAR [8] to each population individually. The T2D data sets have five independent loci with at least one significant variant. The Y-axis is the size of the causal set for each locus. Each dot indicates the causal set for one locus. The violins represent the range of causal set sizes identified by each tool, and the width corresponds to the frequency of that causal set size for that tool.

After these filters, five loci remained to be analyzed by trans-ethnic fine mapping. Linkage disequilibrium (LD) matrices were generated from the 1000 Genomes project [17], with “European” and “East Asian” as the population names, using the “CalcLD_1KG_VCF.py” script from the PAINTOR [20] GitHub repository. As a final step, we pruned groups of SNPs that were perfectly correlated with each other in both studies (arbitrarily picking one SNP in the group to retain), as they provided identical information and would cause the LD matrix to be low-rank. The resulting five loci had 9, 10, 10, 42, and 42 SNPs.

We ran CAVIAR [8], the trans-ethnic mode of PAINTOR [11], and MsCAVIAR on these loci, and evaluated their causal set sizes, since these methods have been shown to be well-calibrated and no ground truth is available (Figure 4). For MsCAVIAR, we set the heterogeneity parameter τ^2 (Methods) to 0.5. For CAVIAR, we evaluated its performance when applying it to only the Asian (BBJ) data or to only the European (UKB) data. For all methods, we set the posterior probability threshold ρ^* to 80% and set the maximum number of causal SNPs to 3.

Whereas the original five loci contained a combined 113 SNPs, MsCAVIAR yielded causal set sizes of 3, 4, 4, 5, and 6, for a total of 22 SNPs (80.53% reduction). CAVIAR on BBJ returned 23 SNPs, CAVIAR on UKB returned 35 SNPs, and Trans-ethnic PAINTOR returned 28 SNPs. Thus, MsCAVIAR yielded a reduction in combined set size of 4.34%, 37.14%, and 21.34%, respectively, compared to those alternative approaches. The Wilcoxon signed-rank test p-values for the alternate hypothesis that MsCAVIAR’s returned causal set sizes were smaller than those given by CAVIAR on BBJ, CAVIAR on UKB, and trans-ethnic PAINTOR were 0.159, 0.019, and 0.031, respectively. All methods ran in less than two minutes for all loci combined, with CAVIAR running the fastest, followed by MsCAVIAR, followed by trans-ethnic PAINTOR.

Based on these findings, the BBJ data seems to have an LD structure that is more favorable to fine mapping for these particular loci for T2D, since CAVIAR on BBJ refined loci much more than CAVIAR on UKB. However, MsCAVIAR was able to leverage both studies to achieve the best resolution. It is interesting that trans-ethnic PAINTOR refined loci less than CAVIAR on BBJ. This may be due to a lack of explicit modeling of heterogeneity and differing sample sizes between

the studies in PAINTOR’s model. Overall, the finding that MsCAVIAR improves fine mapping resolution versus these other approaches is consistent with the findings in our simulation study.

Discussion

In this work, we introduced MsCAVIAR, a method for identifying causal variants in associated regions while leveraging information from multiple studies. Our approach requires only summary statistics as opposed to genotype data and handles heterogeneity of effect sizes, differing sample sizes, and different LD structures between studies, making trans-ethnic fine mapping an ideal application. We demonstrated that our method is well-calibrated in simulation studies and improves fine-mapping resolution in both simulated and real data. MsCAVIAR is available as free and open source software (<https://github.com/nlapier2/MsCAVIAR>).

We make several important assumptions in this model, which may not always be true. While it has been shown that many causal SNPs are shared across populations [13], this may not always be the case. Ideally, this would be obvious from the summary statistics because the population in which the SNP is causal should have much more association signal, in which case one should just apply CAVIAR or a similar single-study method to that study. We also assume that all studies are drawn with equal heterogeneity τ^2 . This is unlikely to be true if multiple studies are from a single population while another study is from a different population. Since the primary benefit of MsCAVIAR is its utilization of varying LD structures, it is unlikely that multiple studies from the same population will confer much benefit. Therefore, we recommend using only one study from each population. However, it is still possible that even ostensibly different populations may be more similar to each other at certain loci than other populations. Therefore, we plan to extend our method to handle this case in future work.

Methods

Fine mapping in a single study

We now describe a standard approach for fine mapping significant variants from a genome-wide association study (GWAS). In the GWAS, let there be n individuals, all of whom have been genotyped at m variants. For each individual j , we measure a quantitative trait y_j , resulting in the $n \times 1$ column vector Y of phenotypic values. We denote G as the $n \times m$ matrix of the genotypes where $g_{ij} \in \{0, 1, 2\}$ is the minor allele count for the j th individual at variant i . We standardize G according to the population proportion p of the minor allele and denote this as X where $x_{ij} \in \left\{ \frac{-2p}{\sqrt{2p(1-p)}}, \frac{1-2p}{\sqrt{2p(1-p)}}, \frac{2-2p}{\sqrt{2p(1-p)}} \right\}$.

We assume Fisher’s polygenic model, which means Y is normally distributed and each variant x_i has a linear effect on Y . We, therefore, have the following model:

$$Y = \mu 1 + \sum_{i=1}^m \beta_i x_i + e \quad (1)$$

where β_i is the effect size variant x_i and e is variation in Y not explained by additive genetic effects. e is an $n \times 1$ vector and follows the Gaussian distribution $e \sim \mathcal{N}(0, \sigma_e I)$ where each individual’s residual is independently and identically distributed. From this linear model, we define $\lambda_i = \frac{\beta_i \sqrt{n_i}}{\sigma_e}$, the standardized effect size of the i th variant, which is also referred to as the “non-centrality parameter” (NCP). Let $\Lambda = [\lambda_1 \dots \lambda_m]$ be an $m \times 1$ column vector of non-centrality parameters. Furthermore, we define the the summary statistic s_i for each effect size β_i , where

$s_i = \frac{\hat{\beta}_i \sqrt{n_i}}{\sigma_e}$, where $s_i = \frac{\hat{\beta}_i \sqrt{n_i}}{\sigma_e} \sim \mathcal{N}(\frac{\beta_i \sqrt{n_i}}{\sigma_e}, 1)$. Let S be an $m \times 1$ vector of summary statistics measured for each variant. As previously shown [8, 21], S follows a multivariate normal (MVN) distribution, $S|\Lambda \sim \mathcal{N}(\Lambda, \Sigma)$ where Σ is the pairwise correlation structure between variants (LD). The expected value of each statistic s_i is a function of its correlation to a causal variant.

While the values of Λ are a function of a variant's relationship to a causal variant, the vector itself does not indicate the variant's causal status; therefore, we introduce an $m \times 1$ binary vector $C = \{0, 1\}^m$ for indicating whether a variant truly does have a non-zero effect size. We will now define Λ_C where each index λ_{C_i} is as follows:

$$\lambda_{C_i} = \begin{cases} 0, & \text{if } c_i = 0 \\ \lambda_i, & \text{if } c_i = 1 \end{cases} \quad (2)$$

This means that if $C = [1 \dots 1]$, $\Lambda_C = \Lambda$ because all variants would be causal. The distribution of Λ_C can be defined as:

$$\Lambda_C|C \sim \mathcal{N}(0, \Sigma_C) \quad (3)$$

where

$$\Sigma_C = \begin{cases} 0, & \text{if } i \neq j. \\ \sigma, & \text{if } i \text{ is causal.} \\ \epsilon, & \text{if } i \text{ is not causal.} \end{cases} \quad (4)$$

and where ϵ is a small constant to ensure that the matrix Σ_C is full rank. Here, and below, we use the shorthand σ to represent the variance of the λ_{C_i} (see the subsection "Extending MsCAVIAR to different sample sizes" for details on this parameter). The off-diagonals of Σ_C are zero because the effect sizes of causal variants are independent of one another.

We will now more formally define Λ as $\Lambda = \Sigma \Lambda_C$, which is to say the non-zero values in Λ are either due to the variant being truly causal or the variant's correlation structure with the causal variant(s). This and the fact that LD structure is symmetric $\Sigma = \Sigma^T$ leads to the following distribution for $\Lambda|C$:

$$(\Lambda|C) \sim \mathcal{N}(0, \Sigma \Sigma_C \Sigma) \quad (5)$$

We will now define γ as the probability of a variant being causal, which makes the causal status for the i th variant a Bernoulli random variable with the following probability mass function: $f(c_i; \gamma) = \gamma^{c_i} (1 - \gamma)^{1-c_i}$. We assume the causal status for each variant is independent of the other variants, leading to the following prior for the our indicator vector: $P(C) = \prod_{j=1}^m \gamma^{c_j} (1 - \gamma)^{1-c_j}$. Assuming that each variant has a probability γ of having a causal effect, the prior can then be written as follows:

$$P(\Lambda, C) = P(\Lambda|C)P(C) = f(\Lambda, 0, \Sigma_C) \prod_{j=1}^m \gamma^{c_j} (1 - \gamma)^{1-c_j} \quad (6)$$

where $f(\Lambda, 0, \Sigma_C)$ is the probability density function shown in equation 5.

We determine which variants are causal by calculating the posterior probability of each configuration $C^* \in \mathcal{C}$, where \mathcal{C} is the set of all possible configurations, given the set of summary statistics:

$$P(C^*|S) = \frac{P(S|C^*)P(C^*)}{\sum_{c \in \mathcal{C}} P(S|c)P(c)} = \frac{\int_{\Lambda_{C^*}} P(S|\Lambda, C^*)P(\Lambda = \Sigma \Lambda_{C^*}, C^*)d\Lambda_{C^*}}{\sum_{c \in \mathcal{C}} \int_{\Lambda_c} P(S|\Lambda, c)P(\Lambda = \Sigma \Lambda_c, c)d\Lambda_c} \quad (7)$$

For us to calculate the posterior probability of C^* given S , we need to integrate over all possible values for the non-centrality parameters of the causal variants in Λ in order to get the values of Λ that makes observing S most probable.

Given a set of \mathcal{K} SNPs, $\mathcal{C}_{\mathcal{K}}$, the posterior probability that this set of SNPs contains all the causal SNPs can then be calculated as follows:

$$P(\mathcal{C}_{\mathcal{K}}|S) = \sum_{C^* \in \mathcal{C}_{\mathcal{K}}} P(C^*, \Lambda_{C^*}|S)$$

The goal is then to find the minimum-sized set \mathcal{K}^* that has a posterior probability of at least ρ^* , called the “ ρ^* confidence set”:

$$P(\mathcal{C}_{\mathcal{K}^*}|S) \geq \rho^*$$

This is done by evaluating causal configuration vectors with only one non-zero element, and then those with two non-zero elements, and so on until the end condition above is met.

Efficient computation of likelihood functions

The integral above is intractable. Fortunately, a closed-form solution is available due to the fact that, when a conjugate prior is multivariate normally distributed, its posterior predictive distribution is also multivariate normal. As shown above, $S|\Lambda \sim \mathcal{N}(\Lambda, \Sigma)$ and $(\Lambda|C) \sim \mathcal{N}(0, \Sigma \Sigma_C \Sigma)$. The posterior predictive form of S is then

$$S \sim \mathcal{N}(0, \Sigma + \Sigma \Sigma_C \Sigma) \quad (8)$$

However, computing the likelihood of S with this distribution is still computationally expensive. Consider the multivariate normal probability density function, assuming the variable Z below is MVN distributed with mean μ and covariance matrix Σ :

$$f(Z; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2}(Z - \mu)^T \Sigma^{-1}(Z - \mu)\right)$$

For S , the covariance matrix is $\Sigma + \Sigma \Sigma_C \Sigma$, which has dimension $(M \times M)$. Taking the determinant or inverse of this covariance matrix, as required by the above likelihood function, would take $O(m^3)$ time. Here, we demonstrate how to compute this likelihood efficiently, leveraging insights from several studies that have explored this topic [9, 10, 22].

We need to compute $S^T(\Sigma + \Sigma \Sigma_C \Sigma)^{-1}S$ and $|\Sigma + \Sigma \Sigma_C \Sigma|$ (note that our μ is 0). We can factor out Σ from both of the equations above:

$$S^T(\Sigma + \Sigma \Sigma_C \Sigma)^{-1}S = S^T \Sigma^{-1}(I + \Sigma_C \Sigma)^{-1}S$$

$$|\Sigma + \Sigma \Sigma_C \Sigma| = |\Sigma| |I + \Sigma_C \Sigma|$$

Notably, $S^T \Sigma^{-1}$ and $|\Sigma|$ can be computed once and re-used for every causal configuration Σ_C . Below, we assume Σ is of full-rank; Lozano et. al [22] show how to address the low-rank case.

We use the Woodbury matrix identity [23], below, to speed up the matrix inversion equation:

$$(A + UEV)^{-1} = A^{-1} - A^{-1}U(E^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Here, we set $A = I_{m \times m}$, $E = I_{k \times k}$, and $UV = \Sigma_C \Sigma$. In particular, U is the $(m \times k)$ matrix of rows corresponding to causal SNPs in Σ_C . We are taking advantage of the fact that rows corresponding to non-causal SNPs are zeros and thus do not affect the matrix multiplication. Similarly, V is the corresponding columns of Σ , and is $(k \times m)$. Applying the Woodbury matrix identity to our case, we get:

$$\begin{aligned} (I_{m \times m} + \Sigma_C \Sigma)^{-1} &= (I_{m \times m} + UV)^{-1} \\ &= I_{m \times m}^{-1} - I_{m \times m}^{-1}U(I_{k \times k}^{-1} + VI_{k \times k}^{-1}U)^{-1}VI_{m \times m}^{-1} \\ &= I_{m \times m} - U(I_{k \times k} + VU)^{-1}V \end{aligned} \quad (9)$$

Crucially, we are now inverting a $(k \times k)$ matrix instead of an $(m \times m)$ matrix, where $k \ll m$ since most SNPs are not causal [22]. We use Sylvester's determinant identity [24] to speed up the determinant computation as follows:

$$|I_{m \times m} + UV| = |I_{k \times k} + VU|$$

Similarly, we are computing the determinant of a $(k \times k)$ matrix instead of an $(m \times m)$ matrix. Using these speedups, the computation of the likelihood function of S is reduced from $O(m^3)$ to $O(k^3)$ plus some $O(mk^2)$ matrix multiplication operations, which is tractable under the reasonable assumption that each locus has at most $k = 3$ causal SNPs.

Fine mapping across multiple studies

As GWAS continue to grow in size, frequency, and diversity, there is an increasing need for fine mapping methods that leverage results from multiple studies of the same trait. A simple approach is to assume that there is one true non-centrality parameter for every variant; therefore Λ_C is identical across studies. This approach is referred to as a fixed effects model. In this case, the t th study's $\Lambda_{Ct} = \Lambda_C$.

While there is evidence that many causal SNPs are shared across populations [13], the assumption that the true causal non-centrality vector Λ_C is the same across studies is unrealistic, especially when the studies are measured in different ethnic groups [12, 14, 11].

We relax this assumption by utilizing a random effects model, in which each study t is allowed to have a different Λ_{Ct} . Under this model, a causal SNP q has an overall mean non-centrality parameter, which we denote with the scalar λ_q , from which the non-centrality parameter for SNP q in each study t , denoted by the scalar λ_{qt} , is drawn with heterogeneity (variance) τ^2 . According to the polygenic model, λ_q is distributed as $\lambda_q \sim \mathcal{N}(0, \sigma)$; therefore, λ_{qt} is distributed as $\lambda_{qt} \sim \mathcal{N}(\lambda_q, \tau^2)$. Consequently, the vector Λ_{Cq} for this SNP across all studies will have the following distribution:

$$\Lambda_{Cq} \sim \mathcal{N}(0, \sigma 11^T + \tau^2 I) \quad (10)$$

where T is the number of studies, 1 is a $(T \times T)$ matrix of 1s, and I is the $(T \times T)$ identity matrix. Intuitively, since the SNP q was drawn with variance σ , this variance component is shared across studies, while the variance component τ^2 is study-specific and therefore it is only present along the diagonal of the covariance matrix. If a variant is not causal, its true effect size should be zero. We construct a matrix Λ_C of size $(mT \times mT)$, where m is the number of SNPs and each

row corresponds to the T -length vector Λ_{Cq} corresponding to SNP q . In practice, we ensure that this matrix is full-rank by drawing the non-causal SNPs according to $\Lambda_{Cq} \sim \mathcal{N}(0, \epsilon I)$, where ϵ is a small constant.

From this we will now build out the posterior probability of $P(C^*|S)$ similarly to equation 7. Now instead of $\Lambda_t = \Sigma_t \Lambda_C$ for study t , we have to account for $\Lambda = \Sigma_t \Lambda_{C_t}$ where Λ_t is drawn from a multivariate normal distribution. This means we have to integrate over the domain-space of Λ_{C_t} to as well as Λ_C to describe $P(C^*|S_t) = \frac{P(S_t|C^*)P(C^*)}{\sum_{C \in C} P(S_t|C)P(C)}$

$$P(C^*|S_t) = \frac{\int_{\Lambda_{C_t}^*} P(S_t|\Lambda_t, C^*) \int_{\Lambda_{C^*}} P(\Lambda_t = \Sigma_t \Lambda_{C_t}^* | \Lambda_{C^*}, C^*) P(\Lambda_{C^*}, C^*) d\Lambda_{C^*} d\Lambda_{C_t}^*}{\sum_{c \in C} P(S|\Lambda, c) \int_{\Lambda_{c_t}} P(S_t|\Lambda_t, c) \int_{\Lambda_c} P(\Lambda_t = \Sigma_t \Lambda_{c_t} | \Lambda_c, c) P(\Lambda_c, c) d\Lambda_c d\Lambda_{c_t}} \quad (11)$$

Efficient meta-analysis

Now that we have described the distribution of each SNP in our meta-analysis, we show how to jointly analyze them. We begin by explicitly defining the structure of the covariance matrix between studies by way of a small example with three SNPs at a locus in two different studies. Since the covariance of a matrix is undefined, we denote $\text{vec}(\Lambda_C)$ as the vectorized form of the original matrix (Λ_C). Concretely:

$$\text{vec}(\Lambda_C) = \text{vec} \left(\begin{bmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \\ \lambda_{13} & \lambda_{23} \end{bmatrix} \right) = \begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{21} \\ \lambda_{22} \\ \lambda_{23} \end{bmatrix}$$

Assume SNPs 1 and 3 are causal and SNP 2 is not causal. Then the vectorized form of the non-centrality parameters given the causal statuses has the following multivariate normal distribution:

$$(\text{vec}(\Lambda_C) | \text{vec}(C)) \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma + \tau^2 & 0 & 0 & \sigma & 0 & 0 \\ 0 & \epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma + \tau^2 & 0 & 0 & \sigma \\ \sigma & 0 & 0 & \sigma + \tau^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \epsilon & 0 \\ 0 & 0 & \sigma & 0 & 0 & \sigma + \tau^2 \end{bmatrix} \right)$$

We call the covariance matrix above Σ_C . Viewing Σ_C as having a block structure, the blocks along the diagonal represent SNPs from the same study, while off-diagonal blocks represent SNPs from different studies. Here Σ_C is $(3 \times 2 \times 3 \times 2) = (6 \times 6)$; in general, for m SNPs and n studies, Σ_C will be $(mn \times mn)$. In other words, there will be an $(n \times n)$ grid of $(m \times m)$ blocks. Within each block, the diagonal represents each SNP's variance, while the off-diagonal represents covariation between different SNPs. As SNPs are assumed to be independent, these are always 0. There are two variance components: the global genetic variance σ from which the global mean non-centrality parameter for a SNP is drawn, and the heterogeneity between studies τ^2 . When a SNP is causal,

its variance (its covariance with itself in the same study) will contain both variance components ($\tau^2 + \sigma$), while its covariance with the same SNP in a different study will be σ , because they were drawn from the same overall non-centrality parameter with variance σ but were drawn separately with variance τ^2 .

The Σ_C above, leaving aside ϵ for now, can alternately be written in the more-compact form

$$\Sigma_C = \begin{bmatrix} \tau^2 + \sigma & \sigma \\ \sigma & \tau^2 + \sigma \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where \otimes represents the Kronecker product operator. This can be further condensed and generalized into:

$$\Sigma_C = (\tau^2 I_n + \sigma 1_n 1_n^T) \otimes \text{diag}(1_{\text{causal}})_m$$

where n is the number of studies, m is the number of SNPs, $1_n 1_n^T$ is the $(n \times n)$ matrix of all 1s, I_n is the $(n \times n)$ identity matrix, and $\text{diag}(1_{\text{causal}})_m$ is an $(m \times m)$ diagonal matrix whose diagonal entries are given by the $(1 \times m)$ indicator vector 1_{causal} whose entries i are 1 if SNP i is causal and 0 otherwise.

As with CAVIAR, the ϵ entries along the diagonal are small numbers to ensure full rank. Also note that the CAVIAR model is a specific case of this model, in which there is only one study and thus there is no τ^2 component. The CAVIAR Σ_C has the same structure as the upper left block in the Σ_C above, when there are 3 SNPs and τ^2 is set to 0.

The efficient computation properties for the single-study case also apply to the multiple-study case. In the latter setting, the matrices that need to be inverted are $(mn \times mn)$ instead of $(m \times m)$, where m and n are the number of SNPs in a locus and the number of studies, respectively. Consequently, in the Woodbury matrix identity equations, U and V are $(mn \times kn)$ and $(kn \times mn)$, respectively, where $k \ll m$ is the number of causal SNPs, and the matrix given by the Woodbury identity is $(kn \times kn)$. Sylvester's determinant identity gives a matrix of this size as well. The computation time is thus reduced from $(mn \times mn)$ to $(kn \times kn)$.

Extending MsCAVIAR to different sample sizes

Previously, we have assumed the i th SNP has one true mean non-centrality parameter, λ_i . This simplifying assumption implies that all studies have the same sample size n , as the non-centrality parameter is a function of sample size $\lambda_i = \frac{\beta_i \sqrt{n}}{\sigma_e}$. We are able to relax this assumption in MsCAVIAR to accommodate studies with differing sample sizes and in doing so will need to more precisely define σ . Additionally, we can no longer assume the summary statistic $s_i = \frac{\hat{\beta}_i \sqrt{n}}{\sigma_e}$ is drawn according to the global mean λ_i ; instead we will need to assume that study m 's observed effect size $\hat{\beta}_{m,i}$ is an unbiased estimate of the true effect size of the SNP, β_i , across studies where

$$\hat{\beta}_{m,i} \sim \mathcal{N}(\beta_i, \frac{\tau^2 \sigma_e^2}{\sqrt{n_m}})$$

We now draw the true effect size of the i th SNP according to $\beta_i \sim \mathcal{N}(0, \sigma_g^2)$; therefore, the mean effect size is 0 and the variance in effect size is the variance explained by this variant under an additive model, which we denote with σ_g^2 . We will again draw the m th study's non-centrality

parameter for variant i according to this model. Each study m has its own sample size n_m , environmental component σ_{e_m} , and we draw it with heterogeneity parameter τ^2 as previously defined, so

$$\lambda_{m,i} \sim \mathcal{N}\left(\frac{\beta_i}{\sigma_{e_m}}\sqrt{n_m}, \tau^2\right)$$

We will now operate under the standard assumption that σ_e has been standardized ($\sigma_e = 1$), so

$$\lambda_{m,i} \sim \mathcal{N}(\beta_i\sqrt{n_m}, \tau^2)$$

Using our previous definition for a single study, we now have

$$\Lambda_C|C \sim \mathcal{N}(0, \Sigma_C) \quad (12)$$

where

$$\Sigma_C = \begin{cases} 0, & \text{if } i \neq j. \\ \sigma, & \text{if } i \text{ is causal.} \\ \epsilon, & \text{if } i \text{ is not causal.} \end{cases} \quad (13)$$

We now define σ more formally to be $\sigma_g n_m$ for the m th study. When we consider our matrix

$$\Sigma_C = \begin{bmatrix} \tau^2 + \sigma & \sigma \\ \sigma & \tau^2 + \sigma \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The σ along the diagonal is defined identically to the precise single study definition; however, when modeling multiple studies, this adjustment changes the covariance between causal variant for two studies. We now define $\sigma = \sqrt{n_x}\sqrt{n_y}\sigma_g^2$ for two studies x and y with population sizes n_x and n_y . Note that if two studies have the same population size n , we get the original definition of $\sigma = \sqrt{n}\sqrt{n}\sigma_g^2 = n\sigma_g^2$ (recalling that here we are assuming $\sigma_e^2 = 1$).

Parameter setting in practice

Traditionally, the effect size $\beta \sim \mathcal{N}(0, \sigma_g^2)$ would be derived as a notion of the per-snp heritability. Here we do not define σ_g^2 as such, but rather treat it as an abstraction: we avoid making any assumptions on how heritable the given trait is and how that heritability is partitioned between loci. The way we set this parameter in practice is as a parameter for statistical power. If study m has the smallest sample size, we set this value such that $\sigma_g^2 n_m = 5.2$ for all variants. Then the NCP for variant i in the corresponding study m is $\lambda_{m,i} \sim \mathcal{N}(5.2, \tau^2)$. For another study x with larger sample size, its NCP is drawn as $\lambda_{x,i} \sim \mathcal{N}(5.2\sqrt{\frac{n_x}{n_m}}, \tau^2)$. This value of σ_g^2 may not represent the actual heritability partitioning, but we set the parameter this way in our method for the practical purpose of giving MsCAVIAR power to fine map borderline significant variants in the smallest study.

References

1. M. Ikeda, A. Takahashi, Y. Kamatani, Y. Okahisa, H. Kunugi, N. Mori, T. Sasaki, T. Ohmori, Y. Okamoto, H. Kawasaki, *et al.*, "A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder," *Molecular psychiatry*, vol. 23, no. 3, p. 639, 2018.
2. L. G. Fritsche, W. Igl, J. N. C. Bailey, F. Grassmann, S. Sengupta, J. L. Bragg-Gresham, K. P. Burdon, S. J. Hebbaring, C. Wen, M. Gorski, *et al.*, "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants," *Nature genetics*, vol. 48, no. 2, p. 134, 2016.
3. A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, *et al.*, "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, p. 197, 2015.
4. D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping," *Nature reviews. Genetics*, 2018.
5. L. L. Faye, M. J. Machiela, P. Kraft, S. B. Bull, and L. Sun, "Re-ranking sequencing variants in the post-gwas era for accurate causal variant identification," *PLoS genetics*, vol. 9, no. 8, p. e1003609, 2013.
6. J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, J. M. Howson, A. Auton, S. Myers, A. Morris, *et al.*, "Bayesian refinement of association signals for 14 loci in 3 common diseases," *Nature genetics*, vol. 44, no. 12, p. 1294, 2012.
7. A. H. Beecham, N. A. Patsopoulos, D. K. Xifara, M. F. Davis, A. Kempainen, C. Cotsapas, T. S. Shah, C. Spencer, D. Booth, A. Goris, *et al.*, "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis," *Nature genetics*, vol. 45, no. 11, p. 1353, 2013.
8. F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin, "Identifying causal variants at loci with multiple signals of association," *Genetics*, pp. genetics-114, 2014.
9. W. Chen, B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland, and D. J. Schaid, "Fine mapping causal variants with an approximate bayesian method using marginal test statistics," *Genetics*, vol. 200, no. 3, pp. 719-736, 2015.
10. C. Benner, C. C. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen, "Finemap: efficient variable selection using summary data from genome-wide association studies," *Bioinformatics*, vol. 32, no. 10, pp. 1493-1501, 2016.
11. G. Kichaev and B. Pasaniuc, "Leveraging functional-annotation data in trans-ethnic fine-mapping studies," *The American Journal of Human Genetics*, vol. 97, no. 2, pp. 260-271, 2015.
12. R. Mägi, M. Horikoshi, T. Sofer, A. Mahajan, H. Kitajima, N. Franceschini, M. I. McCarthy, T.-G. C. COGENT-Kidney Consortium, and A. P. Morris, "Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution," *Human molecular genetics*, vol. 26, no. 18, pp. 3639-3650, 2017.
13. U. M. Marigorta and A. Navarro, "High trans-ethnic replicability of gwas results implies common causal variants," *PLoS genetics*, vol. 9, no. 6, p. e1003566, 2013.
14. A. P. Morris, "Transethnic meta-analysis of genomewide association studies," *Genetic epidemiology*, vol. 35, no. 8, pp. 809-822, 2011.
15. B. Han and E. Eskin, "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies," *The American Journal of Human Genetics*, vol. 88, no. 5, pp. 586-598, 2011.
16. Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, *et al.*, "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature*, vol. 506, no. 7488, p. 376, 2014.
17. . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.
18. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.
19. K. Suzuki, M. Akiyama, K. Ishigaki, M. Kanai, J. Hosoe, N. Shojima, A. Hozawa, A. Kadota, K. Kuriki, M. Naito, *et al.*, "Identification of 28 new susceptibility loci for type 2 diabetes in the japanese population," *Nature genetics*, vol. 51, no. 3, p. 379, 2019.
20. G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc, "Integrating functional data to prioritize causal variants in statistical fine-mapping studies," *PLoS genetics*, vol. 10, no. 10, p. e1004722, 2014.
21. B. Han, H. M. Kang, and E. Eskin, "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers," *PLoS genetics*, vol. 5, no. 4, p. e1000456, 2009.
22. J. A. Lozano, F. Hormozdiari, J. W. J. Joo, B. Han, and E. Eskin, "The multivariate normal distribution framework for analyzing association studies," *bioRxiv*, p. 208199, 2017.

23. H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *Siam Review*, vol. 23, no. 1, pp. 53–60, 1981.
24. A. G. Akritas, E. K. Akritas, and G. I. Malaschonok, "Various proofs of sylvester's (determinant) identity," *Mathematics and Computers in Simulation*, vol. 42, no. 4-6, pp. 585–593, 1996.